TITLE

HPC-oriented computer aided drug design approaches in the exascale era

KEYWORDS

Artificial intelligence, computer aided drug design, exascale computing, high performance computing, molecular simulations

ABSTRACT

Introduction

In 2023 the first exascale supercomputer was opened to the public in the US. With a demonstrated 1.1 exaflops of performance, Frontier represents an unprecedented breakthrough in high performance computing (HPC). Currently, more (and more powerful) machines are being installed worldwide. Computer-aided drug design (CADD) is one of the fields of computational science that can greatly benefit from exascale computing for the benefit of the whole society. However, scaling CADD approaches to exploit exascale machines requires new algorithmic and software solutions.

Areas covered

Here, the authors consider physics-based and machine learning(ML)-aided techniques for the design of small molecule binders capable of leveraging modern parallel computer architectures. Specifically, the authors focus on HPC-oriented large-scale applications from the past three years that were enabled by (pre)exascale supercomputers by running on tens of thousands of accelerated nodes.

Expert opinion

In the area of ML, exascale computers can enable the training of generative models with unprecedented predictive power to design novel ligands, provided large amounts of high-quality data are available. Exascale computers could also unlock the potential of accurate ML-aided physics-based methods to boost the success rate of structure based drug design campaigns. Currently, however, methodological developments are still required to allow routine large-scale applications of such rigorous approaches.

ARTICLE HIGHLIGHTS BOX

- Recent studies pushed the boundaries of HPC-oriented physics-based and machine learning-aided computational tools for the discovery and development of small molecule binders
- In machine learning-based drug design, exascale machines can enable fast training of generative models on massive data sets to guide the design of new ligands
- In structure based drug design, exascale machine can enable routine docking of billions of small molecules, as well as the use of accurate molecular dynamics-based alchemical free energy methods via scalable automated workflows
- Rigorous quantum mechanics/molecular mechanics molecular dynamics simulations enabled by extremely scalable software can increase hit to lead success rate of virtual screening campaigns by providing high quality structural information
- Exascale computing will have a transformative impact in computer aided drug design only by increasing the throughput of these (often still time consuming) HPC-oriented approaches: data-driven methods provide viable solutions, however, further methodological developments are still required

1. Introduction

The process of drug discovery is one of the most difficult, time-consuming and resource-intensive endeavors in the healthcare industry. Depending on the therapy area, the cost of getting a novel drug to the market is estimated to range between 0.3 to 4.5 billion USD and 5 to 15 years of development[1,2,3,4,5].

After the identification of a promising biological target (e.g., a protein), the initial stage of drug design campaigns typically involves two main tasks: i) the identification of chemical compounds with promising activity (hit molecules); ii) the optimization of the compounds properties (e.g., potency, selectivity, toxicity, pharmacokinetics) in the hit-to-lead and lead optimization stages to generate suitable candidates for pre-clinical and clinical trials. Performing these preliminary experiments *in silico* can drastically reduce the costs and the research time for developing a new drug molecule [6].

Computer aided drug design (CADD) is an umbrella term indicating the collection of computational techniques used to discover, develop, and optimize chemical compounds. These methods include docking-based virtual screening, molecular modeling, and quantitative structure-activity relationships (QSAR)[7,8,9]. More recently, these have been flanked by artificial intelligence methods based on neural networks (NNs) [8,10]. The goal of such computational tools is to identify promising candidates, whose efficacy is subsequently tested *in vitro* and, eventually, *in vivo*. Nowadays, CADD is a well-established tool used by pharmaceutical companies and academic institutions to accelerate the early stages of drug

design[9,11,12]. Indeed, CADD has already played an important role in discovering many drugs that reached the consumer market[13,14,15,16,17].

Traditionally, CADD approaches are broadly classified into structure based drug design (SBDD) and ligand based drug design (LBDD). SBDD methods explicitly incorporate the information on the 3D structure of the biomolecular target – usually obtained through X-ray crystallography[18], nuclear magnetic resonance spectroscopy[19], or electron microscopy[20]. This enables the use of (i) docking methodologies to perform the virtual screening of large databases of small molecules to obtain a set of possible hits and an estimate of their binding mode, as well as (ii) the use of accurate atomistic simulations (e.g, alchemical techniques) to rank compounds based on their binding affinities. These approaches have found widespread use in the industry[21,22]. By including the 3D structure of the target, SBDD is expected to improve the compounds' specificity and to provide valuable insights into the relevant protein-ligand interactions, which facilitates the rational design of more effective drugs. It is interesting to note that with the advent of accurate deep learning models for protein structure prediction [23,24], nowadays fully *in-silico* SBDD campaigns can be envisioned.

In cases where no reliable 3D structures of the targets are available, LBDD approaches can be used. In this case, the starting point is usually a set of known hits to the specific target, usually determined experimentally via (expensive and time consuming) high throughput biochemical assays [25,26]. LBDD methods can then be used to screen large databases of small molecules to search for compounds with similar chemical-physical properties [27].

Two critical aspects that affect the success rate of the above mentioned SBDD and LBDD approaches are the size of the searchable chemical space of drug-like compounds and the accuracy of the models used for the predictions of their properties. As is typical in these

scenarios, there is often a trade-off between computational cost and accuracy of the methodologies, which limits their domain of application. For example, docking simulations [28] can quickly explore fairly large datasets of small molecule binders with simplified empirical models, but their resolution is not sufficient to be employed in binding optimization, where the differences between tested molecules is small. On the other hand, atomistic molecular dynamics (MD)-based free energy methods are routinely applied to test congeneric series, but their computational cost limits them to at most a few hundreds of compounds [21] or even only a handful of them when employing quantum-mechanically accurate simulations [29,30].

More recently there has been a lot of excitement about the capability of deep learning techniques to provide accurate predictions at much lower computational cost [31]. However, training these models is extremely computation and data-hungry and testing a single new idea typically requires days to weeks on large parallel machines.

The unique parallelization capabilities offered by exascale machines have the potential to considerably push forward the limits of these trade-offs, enabling an efficient exploration of the chemical space using more accurate (physics-based or data-driven) methodologies. This will translate into a reduced time-to-solution and (consequently) cost of the drug discovery process, which impacts patients in the form of, e.g., insurance premiums and taxes.

2. Challenges of exascale computing

Modern supercomputers are extremely complex machines where computational resources are distributed across thousands of compute nodes interconnected via a high throughput and low latency communication network[32]. Each compute node typically consists of a few (typically one or two) central processing units (CPUs) mounted on the same board. CPUs, in turn, can

contain up to few tens of cores residing on the same die (see Figure 1a-c). The cores are the elementary processing elements that execute tasks independently from each other.

An important aspect in HPC is the overhead associated with the fetching of the data from the computer memory. In fact, the memory bandwidth is the main bottleneck limiting the performance of modern computers [33]. In order to reduce this overhead, modern CPU architectures have hierarchical caching mechanisms implemented on chip to store temporary data physically close to each compute unit. Depending on the cache level, usually around 1-10 clock cycles are needed to fetch the data. This has to be compared to hundreds of clock cycles in case of main memory access. It is clear that caching of memory accesses is extremely important to achieve the highest level of performance as cache misses can lead to the loss of hundreds of clock cycles spent retrieving data from the main memory, which can potentially slow down the execution of the program by up to few orders of magnitude. It is the software developer's job to design algorithms that maximize data re-usage on local caches.

Exascale supercomputers push the complexity of distributed architecture to a whole new level by connecting several thousands of *heterogeneous* computing nodes that combine traditional general-purpose CPUs with powerful graphics processing units (GPUs) (see Figure 1d,e). The latter can be thought of as a special type of processing unit hosting hundreds of cores that enables the implementation of massively parallel shared-memory algorithms using dedicated programming models.

An ideal HPC application is one that is able to scale efficiently on all the available resources. This means running on all the nodes while keeping the performance (flops) as close as possible to the nominal value guaranteed by the hardware architecture. Optimization of an application's performance targeting distributed heterogeneous parallel architectures is a very challenging

problem and designing algorithms that can effectively exploit exascale supercomputers requires careful considerations of the hardware organization at all levels, from the single CPU / GPU level up to the multiple nodes level.

3. Opportunities of exascale computing for HPC-oriented CADD

The extreme concurrency provided by exascale supercomputers represents a great opportunity to push the state-of-the art of HPC-oriented CADD for the design of small molecule binders (see Figure 2). Here, we highlight selected contributions that were able to exploit efficiently the unique capabilities of (pre)exascale machines. Specifically, we consider (1) machine learning (ML) methods used to train neuronal networks on unprecedented massive datasets. Exploiting the predictive power of large generative models trained on accurate and large experimental datasets bear the potential to reduce the time to identify in a reliable way novel, candidate antiviral drugs. (2) Docking-based virtual screening of large databases of drug-like molecules, also aided by ML, where exascale computing can be leveraged to expand the size of the explorable chemical space by orders of magnitudes. (3) Highly parallelizable MD-based alchemical free energy methods for binding affinity predictions. Designing automated workflows able to handle thousands of independent MD tasks on exascale computers can boost tremendously the throughput of these accurate methods making them feasible within large-scale drug screening campaigns. (4) Quantum mechanics-based MD simulations, enabled by HPC-oriented codes targeting modern heterogeneous architectures, which can allow routine applications of these rigorous approaches in SBDD.

Most selected publications reviewed here participated in recent editions of the Gordon Bell Special Prize for High Performance Computing-Based COVID-19 Research.

3.1 Training of generative neural network models on large datasets

Applications of neural networks into the drug discovery process have been envisioned since the emergence of deep learning [34]. NNs can learn patterns and correlations from data and generate predictions that can be tested in the laboratory [35]. In recent years, in particular, generative models gathered attention in CADD for their ability to propose novel compounds satisfying user-provided properties [31]. This effort was also spurred by the success of generative models in other fields, which was driven by a dramatic increase in the size of training datasets (hundreds of billions of data points) and models (billions to trillions parameters) [36]. Training NNs on this scale requires vast amounts of power and sophisticated parallelization techniques to spread the workload and the data over computational units [37]. Exascale computing could in principle further push the limits of these techniques. In the context of drug discovery, however, generating billions of high-quality data points through biochemical or physical assays is currently unfeasibly expensive, and thus data often represents a fundamental bottleneck for the adoption of these approaches.

Generative models for small molecules have emerged as a promising route towards this goal since they can be trained in an unsupervised manner on large collections of chemical structures (e.g., in SMILES format) without the need of labeling the compounds with physico-chemical properties. These methods leverage this information to build general and rich representations of molecules that can be later used in specific prediction tasks (38), often after fine tuning the model in a separate stage using smaller datasets.

Jacobs et al. were able to train an autoencoder generative model on >1.6 billion compounds in only 23 minutes, compared to previous state of the art that took a day on 1 million

compounds[39]. This result could be achieved by implementing a parallel training approach leveraging multiple and asynchronously coordinated trainers, which was able to strong scale to the whole Sierra supercomputer at the Lawrence Livermore National Laboratory (LLNL) with close to 100% parallel efficiency [40].

In another recent work, Blanchard et al. pre-trained a transformer-based language model on ~9.6 billion molecules on the Summit supercomputer at Oak-Ridge Leadership Facilities (OLCF) [41] by exploiting massive data parallelism through optimizers capable of handling extremely large batch sizes (more than a million molecules) while limiting overfitting effects [42]. The model was then fine-tuned using a set of thousands of protein targets with binding affinity data [41] combined with a genetic algorithm approach to generate and score drug candidates as inhibitors targeting SARS-CoV-2 Mpro and PLpro proteins. A remarkable result is the reduction of pre-training time from days to hours while using a dataset nearly one order of magnitude larger compared to previous works on the same computer architecture.

3.2 Virtual screening of massive molecular libraries

As the task of virtual screening can be trivially parallelized over compounds, access to exascale machines can enable the exploration of a much larger portion of the drug-like chemical space. In this direction, much progress has been made in the past few years for the task of docking small molecules to their therapeutic target, pushing the boundaries of state-of-the-art SBDD.

In their work, LeGrand and co-authors present the porting, optimization, and validation of the AutoDock-GPU program for large-scale protein-ligand docking calculations on the Summit supercomputer at OLCF [43]. The method was successfully applied to the initial screening of compounds targeting the SARS-CoV-2 virus' Mpro and PLpro proteins. This contribution

represents the first effort to redesign a well-established and widely used docking code targeting modern accelerated HPC infrastructures.

Building on the work of LeGrand and co-authors, Glaser et al. demonstrated extremely fast screening of massive chemical databases for COVID-19 drug discovery on Summit, achieving an order of magnitude reduction in time-to-solution compared to previous methods, docking over one billion compounds to SARS-CoV-2 Mpro and PLpro protein structures[44]. Their GPU-enabled parallel approach is based on an optimized version of AutoDock-GPU to generate and score binding poses. The workflow demonstrated an incredible average rate of 19,028 compounds docked per second, corresponding to a 50× speedup in time to solution compared to previous CPU-based methods. Re-ranking of the most favorable poses generated by the initial low-resolution docking simulations was subsequently used to complete the screening and suggest hit candidates. Specifically, the authors adopted the RFScore-VS family of machine-learning based scoring functions [45], which use descriptors characterizing the interacting atoms of the ligand and the protein in combination with random forests to predict ligand binding affinities [46]. This step of the pipeline adopted parallel database methods on GPUs for analyzing the massive docking output dataset and achieved a further 10× speedup compared to previous methods.

Another notable contribution is the "ab-initio" docking method of the TwoFold code developed by Hsu et al.[47]. This scalable software stack – which builds upon, and improves technical solutions first introduced by AlphaFold2 [48]– not only is able to predict how strongly a drug molecule will bind to a pathogen, but it is also trained to predict the structure of a given protein/ligand complex starting from the protein amino acid sequence and ligand chemical formulas, encoded as SMILES [49]. In TwoFold, the binding poses are predicted by a deep

learning model, while the binding affinity is predicted by the method of the neural tangent kernel (NKL) [50]. Training of the neural network was performed on the Summit supercomputer, using an experimental structural dataset of protein-ligand complexes extracted from a subset of the Protein Data Bank, while the NKL model was trained to predict binding affinities using ~1,4M sequence-ligand pairs and their corresponding experimental IC_{50} values. Training and validating the NKL model involved solving a set of >1M linear equations, which was done on the Frontier exascale supercomputer. The TwoFold approach was tested on a set of 195 protein-ligand complexes matching state-of-the-art implementations in quality and efficiency for binding affinity predictions, while additionally reconstructing protein structures.

3.3 Scalable statistical mechanics-based methods for absolute binding affinities

Molecular dynamics (MD) simulations are a powerful tool that can provide detailed insights on the mechanisms and energetics of molecular recognition phenomena. Compared to standard docking approaches, MD simulations are lower in throughput but thoroughly account for the dynamics of the system[51]. Because of this, besides providing detailed structural information, atomistic MD simulations can in principle provide more accurate predictions for the thermodynamic (and kinetic) binding parameters used to rank small molecules and guide rational drug design[52,53,54].

Among these methods, alchemical free energy perturbation approaches are particularly suited to exploit massively parallel architectures. Within these frameworks, differences between the absolute binding free energies of small molecules binding to a target can be computed via independent or loosely dependent simulations at different points of a suitably defined thermodynamic cycle [55,56]. Simulations for different molecules and thermodynamic points are

easily parallelized as independent jobs on a supercomputer. Exascale computers can considerably reduce time-to-solution by allowing applications to larger sets of protein/ligand complexes, while significantly accelerating each MD task via GPU offloading [57,58,59,60]. The main challenge to scaling these methods stems from the complex setup of the simulations, which require designing automated workflows able to control the status of thousands of tasks and program new MD runs, when required.

The work of Gapsys et al. [61] employed a non-equilibrium alchemical free energy perturbation method[55] to estimate the relative binding affinities for a set of previously published protein-ligand complexes[62] within a highly parallelized workflow. The pipeline made use of local resources for the preparatory stages – having relatively small computer cost – and a cluster for the final stage that involves carrying out MD simulations until satisfactory statistical convergence of binding affinities is achieved. By exploiting job parallelization, a cumulative 200 μs of simulation was generated in only three days to obtain converged relative binding affinities for more than 500 protein-ligand pairs. The work demonstrates how high-throughput prediction of protein-ligand binding affinities is readily achievable with the high accuracy of all-atom alchemical free energy methods, provided that sufficient computational resources are available.

In the work of Li et al.[63], a scalable workflow implementing a free energy perturbation method[64] to compute absolute binding affinities was used to speed up the discovery of antiviral drugs targeting SARS-CoV-2 Mpro and TMPRSS2 proteins. The authors introduced several approximations in the calculation of the binding affinity to perform a virtual screening for more than ten thousands protein-ligand binding systems on a new generation of Tianhe supercomputers using a task management tool specifically developed for automating the whole process, which involved more than 500,000 MD tasks. The best scoring ligands were tested in

further experimental validation in which 50 out of 98 compounds showed significant inhibitory activity towards Mpro, including an inhibitor that showed promising outcomes in subsequent clinical trials.

The embarrassingly parallel nature of alchemical free energy calculations – which do not rely on frequent communications between different tasks – doesn't require the use of supercomputers hosted by HPC centers, but can be efficiently implemented on other extremely parallel computing platforms as well. As a notable example, we mention here the Folding@Home distributed computing platform, which recently reached the exaflop peak performance by running parallel short MD simulations of the SARS-CoV-2 spike protein using computational time on 280,000 GPU and 4.8 million CPU cores donated by the community [65]. This platform was used to run more than 22,000 rigorous alchemical free energy calculations to prioritize the compounds for synthesis in a fully open source consortium for the development of antivirals targeting SARS-CoV-2 Mpro [66].

3.4 Molecular dynamics simulations including quantum mechanical effects

Modern SBDD requires atomistic modeling of proteins interacting with small molecules that are expected to interfere with their functioning to achieve a desired therapeutic effect. Nowadays, most molecular simulations used in SBDD make use of simplified empirical force fields that do not account for (often essential) quantum mechanical phenomena. One way of overcoming this problem is by leveraging rigorous multiscale quantum mechanics / molecular mechanics (QM/MM) MD simulations[67]. QM/MM MD has already been applied to investigate bond forming/breaking processes of covalent inhibitor binding transition-metal based drugs and to study enzymatic reactions for the design of transition state-analog inhibitors[29,30,68].

QM/MM simulations approaches with relatively high-throughput (mainly point energy calculations and structural optimizations) have also been applied to improve the success rate of virtual screening campaigns by refining the starting geometries for docking, building more accurate charge models, and improving the accuracy of scoring functions[69]. While on one hand, QM/MM simulations have a much higher computer cost than force field-based ones, on the other, QM/MM MD codes scale better with increasing number of processors[70], and exascale computers may finally make them fruitable within the realm of drug design[71].

Recent advances in QM/MM software development demonstrated efficient scaling up to >80 kcores on the pre-exascale JUWELS machine at the Juelich Supercomputing Center studying the IDH1 enzyme, a target for the early diagnosis and treatment of brain cancer[72]. This was made possible by the use of an efficient framework, based on a multiple program multiple data approach, that interfaces existing QM and MM software, minimizing communication and preserving the performance of the QM layer, which ultimately dictates the scaling[73]. The studies of Ragavan et al. on the IDH1 enzyme[72,74], not only showcase the extreme scalability of state-of-the-art QM/MM software, but it also represents a clear example where using QM/MM MD simulations is indispensable to obtain high-quality structures for the screening of small molecules using docking approaches.

4. Conclusions

The advent of the exascale era in supercomputing can trigger exciting developments in computer aided drug design potentially cutting time and costs of drug design campaigns by improving the success rate of the initial *in silico* screening phases. It is foreseen that this goal will be achieved by developing integrated approaches combining extremely scalable implementations of AI-based

methods with accurate physics-based molecular simulations. A demonstration of how this ambitious goal could be achieved is provided by the works considered in this review, which we summarize below and in Table 1.

The works of Jacobs et al.[39] and of Blanchard et al.[41] show how modern supercomputers allow extremely fast training of large ML models using massive databases of small molecules to predict binding affinities and suggest candidate hit molecules, demonstrating the power of ML to automate and accelerate drug design. Exascale machines can also push the boundaries of structure based drug design campaigns based on docking simulations. This can be achieved using HPC-oriented implementations of standard algorithms within scalable workflows that allow screening increasingly larger portions of the chemical space, as exemplified by the works of Le Grand et al.[43] and of Glaser et al.[44]. Furthermore, modern supercomputers enable fast training and fruition of AI-based folding algorithms to implement “ab-initio” docking approaches to predict binding affinities as well as ligand/protein structures starting from minimal information (protein primary structure and the ligand chemical formula)[47].

Force-field based classical molecular dynamics simulations allow implementing theoretically rigorous statistical mechanics-based approaches to predict binding affinities. In this context, exascale machines could finally enable high-throughput molecular dynamics-based screening of large libraries of small molecules (or at least refinement of the output of lower resolution methods) beyond standard docking approaches, as shown by recent works of Gapsys et al.[61] and of Li et al.[63].

Finally, using quantum mechanically accurate molecular dynamics simulations is necessary whenever classical, force-field based methods fail to account for important electronic effects. This is the case of ligands binding to metalloproteins (which are estimated to represent 30% to

50% of all known proteins) and covalent ligands[29,75]. Here, multiscale QM/MM MD simulations provide a route to generate high-quality protein structures as a starting point for docking. These applications can only be enabled by extremely scalable HPC-oriented software[76]. A recent example of the crucial role that QM/MM simulations can play in CADD is provided by the works of Ragavan et al.[72,74] on the IDH1 enzyme as a target for the early diagnosis of brain cancer.

5. Expert Opinion

In the last decades we have witnessed a continuous technological advancement in HPC, which recently reached a new milestone by breaking the exascale limit. In the US, Frontier was launched as the first public exascale supercomputer in 2022, Aurora recently entered the top500 list as the second most powerful machine to break the exascale limit (with a half-scale system), OceanLight and Tianhe-3 are operational in China. In Europe, JUPITER is due to launch in Germany in 2025, and there are plans to install exascale supercomputers in other EU countries as well. Designing, installing and operating these infrastructures represents a significant investment of public resources. Furthermore, the average lifetime of supercomputers in an HPC center is typically five years before they are decommissioned and replaced by the next generation machine. It is therefore of paramount importance that these projects hit the ground running, producing high-impact scientific results. Lots of efforts have been spent in the past years preparing for the exascale, setting up the required infrastructure and, most importantly, developing well in advance lighthouse applications to crack important problems in different scientific domains. Notably, in the biology domain, the design of novel therapeutics, including small ligands, has been identified as one of the most pressing problems to be addressed, as

testified by the commitment of large consortia involved in this technological transformation [77,78,79,80].

The advent of exascale machines has been preceded by general hype in the scientific community. However, at the moment, only a handful of exascale machines are operational, and applications in CADD are therefore very few in number. While the papers reviewed here are remarkable achievements that pushed the boundaries of HPC-oriented CADD methods for the design of small ligands in terms of scalability and time-to-solution, it has yet to be proven that this technology can actually have a transformative effect on the field: that is, leading to the design of novel drugs, quicker.

Considering current trends in machine learning, particularly in the context of generative models [31], one can expect particularly impactful results from this area. Here, exascale supercomputers can enable the training of large models, including billions of parameters, using large experimental data sets, including protein structures as well as annotated libraries of thermodynamic parameters characterizing ligand/protein complexes. Nevertheless, while exascale computers in principle provide this capability, there are still bottlenecks related to inaccessibility, and/or lack of high-quality data, as well as limited interpretability, which restrict the application and affect the performance of such models, and must be therefore addressed first [35,81].

Virtual screening campaigns based on docking is a well-established method in structure based drug design that can straightforwardly take advantage of the concurrency provided by exascale computers to expand the size of the searchable chemical space of drug-like molecules. However, the extent to which massive virtual screening campaigns can improve hit to lead success crucially depend on the quality of the scoring function, as well as on the reliability of the target

protein structures. Therefore, the priority in this field should focus on increasing the throughput of more accurate physics-based methods for the prediction of structural and thermodynamic properties [71]. Establishing efficient automated workflows on exascale machines can boost the throughput of molecular dynamics based approaches, such as alchemical free energy calculations, to improve the hit to lead success when incorporated in virtual screening pipelines. However, also application of these methods are limited due to the approximations adopted to describe the interatomic interactions implemented by classical force-fields. Making rigorous multiscale QM/MM MD simulations a standard for pharmacology will represent a major breakthrough in terms of efficiency in discriminating true from false positives in virtual screening campaigns and, for rational drug design, by providing high-quality structural information and microscopic insights for the engineering of novel ligands. This goal can be achieved only by designing HPC-oriented QM software able to scale on heterogeneous architectures [82,83,84,85,86] in combination with efficient QM/MM interfaces [73,87,88].

While state-of-the-art software already enables routine QM/MM simulation of large biological systems of pharmaceutical relevance [72,74], the performance of QM/MM MD simulations are still very far from enabling their direct application in combination with free energy methods for the prediction of thermodynamic and kinetic parameters. Scalable data-driven methods can provide a route to overcome this issue too, for example, by developing ML-based rigorous free energy perturbation approaches [89,90]. These approaches rely on relatively cheap classical MD simulations to generate an ensemble of protein/ligand configurations. Static QM/MM calculations on the generated poses are then performed in an iterative manner to variationally optimize the estimate of the binding free energies at the QM/MM level. These calculations are easily parallelizable and can therefore straightforwardly exploit massively parallel architectures.

Specifically, exascale machines could be efficiently leveraged by implementing the training within workflows able to generate the data and update the neural network in an unsupervised, automated manner. At the present stage, however, there are still methodological developments required to adapt these approaches to multiscale QM/MM simulations in explicit solvent. Alternatively, another promising direction opened by machine learning is that of neural network potentials. While training robust potentials that can be applied to biological systems is still a challenge, these models are sufficiently expressive to provide QM-like accuracy at a fraction of the cost (that is, without severely affecting the performance of MD simulations) and they have already been shown to scale over thousands of GPUs allowing large scale applications [91,92].

BIBLIOGRAPHY

Papers of special note have been highlighted as either of interest [*] or of considerable interest [**] to readers.

1. Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, et al. Drug repurposing: Progress, challenges and recommendations. Nat Rev Drug Discov. 2018;18(1):41–58.
2. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov. 2010 Mar;9(3):203–14.
3. Wouters OJ, McKee M, Luyten J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. JAMA - J Am Med Assoc. 2020;323(9):844–53.
4. Sertkaya A, Beleche T, Jessup A, Sommers BD. Costs of Drug Development and Research and Development Intensity in the US, 2000-2018. JAMA Netw Open. 2024;7(6):1–13.
5. Schlander M, Hernandez-Villafuerte K, Cheng CY, Mestre-Ferrandiz J, Baumann M. How Much Does It Cost to Research and Develop a New Drug? A Systematic Review and Assessment. PharmacoEconomics. 2021;39(11):1243–69.
6. Kiriiri GK, Njogu PM, Mwangi AN. Exploring different approaches to improve the success of drug discovery and development projects: a review. Future J Pharm Sci. 2020 Dec;6(1):27.
7. Villanueva MT. Virtual screening yields refined GPCR agonists. Nat Rev Drug Discov. 2022 Dec;21(12):879–879.
8. Muratov EN, Bajorath J, Sheridan RP, Tetko IV, Filimonov D, Poroikov V, et al. QSAR without borders. Chem Soc Rev. 2020;49(11):3525–64.
9. Sadybekov AV, Katritch V. Computational approaches streamlining drug discovery. Nature. 2023 Apr;616(7958):673–85.
10. Ghasemi F, Mehridehnavi A, Pérez-Garrido A, Pérez-Sánchez H. Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks. Drug Discov Today.

2018;23(10):1784–90.
11. Brogi S. Computational approaches for drug discovery. Molecules. 2019;24(17):1–6.
12. Schneider G, Clark DE. Automated De Novo Drug Design: Are We Nearly There Yet? Angew Chem Int Ed. 2019 Aug;58(32):10792–803.
13. Clark DE. What has computer-aided molecular design ever done for drug discovery? Expert Opin Drug Discov. 2006 Jul;1(2):103–10.
14. Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: Methods and applications. Nat Rev Drug Discov. 2004;3(11):935–49.
15. Talele T, Khedkar S, Rigby A. Successful Applications of Computer Aided Drug Discovery: Moving Drugs from Concept to the Clinic. Curr Top Med Chem. 2010 Jan;10(1):127–41.
16. Wang X, Song K, Li L, Chen L. Structure-Based Drug Design Strategies and Challenges. Curr Top Med Chem. 2018 Sep;18(12):998–1006.
17. Muratov EN, Amaro R, Andrade CH, Brown N, Ekins S, Fourches D, et al. A critical overview of computational approaches employed for COVID-19 drug discovery. Chem Soc Rev. 2021;50(16):9121–51.
18. Papageorgiou AC, Poudel N, Mattsson J. Protein structure analysis and validation with X-ray crystallography. Methods Mol Biol. 2021;2178:377–404.
19. Hu Y, Cheng K, He L, Zhang X, Jiang B, Jiang L, et al. NMR-Based Methods for Protein Analysis. Anal Chem. 2021 Feb;93(4):1866–79.
20. Renaud JP, Chari A, Ciferri C, Liu WT, Rémigy HW, Stark H, et al. Cryo-EM in drug discovery: Achievements, limitations and prospects. Nat Rev Drug Discov. 2018;17(7):471–92.
21. Schindler CEM, Baumann H, Blum A, Böse D, Buchstaller HP, Burgdorf L, et al. Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects. J Chem Inf Model. 2020 Nov;60(11):5457–74.
22. Muegge I, Hu Y. Recent Advances in Alchemical Binding Free Energy Calculations for Drug Discovery. ACS Med Chem Lett. 2023;14(3):244–50.
23. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021 Aug;373(6557):871–6.
24. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630(8016):493–500.
25. Inglese J, Johnson RL, Simeonov A, Xia M, Zheng W, Austin CP, et al. High-throughput screening assays for the identification of chemical probes. Nat Chem Biol. 2007;3(8):466–79.
26. Blay V, Tolani B, Ho SP, Arkin MR. High-Throughput Screening: today's biochemical and cell-based approaches. Drug Discov Today. 2020 Oct;25(10):1807–21.
27. Jung S, Vatheuer H, Czodrowski P. VSFlow: an open-source ligand-based virtual screening tool. J Cheminformatics. 2023;15(1):1–10.
28. Cerqueira NMFSA, Gesto D, Oliveira EF, Santos-Martins D, Brás NF, Sousa SF, et al. Receptor-based virtual screening protocol for drug discovery. Arch Biochem Biophys. 2015 Sep;582:56–67.
29. Riccardi L, Genna V, De Vivo M. Metal–ligand interactions in drug design. Nat Rev Chem. 2018 Jun;2(7):100–12.
30. Schramm VL. Transition States, Analogues, and Drug Development. ACS Chem Biol. 2013 Jan;8(1):71–81.
31. Anstine DM, Isayev O. Generative Models as an Emerging Paradigm in the Chemical Sciences. J Am Chem Soc. 2023;145(16):8736–50.
32. Hager G, Wellein G. Introduction to High Performance Computing for Scientists and Engineers.
33. McKee WAW, A S. Hitting the memory wall: Implications of the obvious. Comput Archit

News. 1995;23(1):20–4.
34. Lecun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
35. Huang K, Fu T, Gao W, Zhao Y, Roohani Y, Leskovec J, et al. Artificial intelligence foundation for therapeutic science. Nat Chem Biol. 2022;18(10):1033–6.
36. Hudson NC, Pauloski JG, Baughman M, Kamatar A, Sakarvadia M, Ward L, et al. Trillion Parameter AI Serving Infrastructure for Scientific Discovery: A Survey and Vision. In: Proceedings of the IEEE/ACM 10th International Conference on Big Data Computing, Applications and Technologies [Internet]. New York, NY, USA: ACM; 2023. p. 1–10. Available from: https://dl.acm.org/doi/10.1145/3632366.3632396
37. Besta M, Hoefler T. Parallel and Distributed Graph Neural Networks: An In-Depth Concurrency Analysis. IEEE Trans Pattern Anal Mach Intell. 2024 May;46(5):2584–606.
38. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, et al. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. ACS Cent Sci. 2018 Feb;4(2):268–76.
39. Jacobs SA, Moon T, McLoughlin K, Jones D, Hysom D, Ahn DH, et al. Enabling rapid COVID-19 small molecule drug design through scalable deep learning of generative models. Int J High Perform Comput Appl. 2021;35(5):469–82.*

The authors trained a generative model on >1.6 billion compounds in only 23 minutes to enable rapid COVID-19 small molecule drug design, reducing by orders of magnitude the time-to-solution compared to previous state-of-the-art approaches.

40. Jacobs SA, Pearce R, Dryden N, Van Essen B. Towards scalable parallel training of deep neural networks. Proc MLHPC 2017 Mach Learn HPC Environ - Held Conjunction SC 2017 Int Conf High Perform Comput Netw Storage Anal. 2017;
41. Blanchard AE, Gounley J, Bhowmik D, Chandra Shekar M, Lyngaas I, Gao S, et al. Language models for the prediction of SARS-CoV-2 inhibitors. Int J High Perform Comput Appl. 2022;36(5–6):587–602.
42. You Y, Li J, Reddi S, Hseu J, Kumar S, Bhojanapalli S, et al. Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. 8th Int Conf Learn Represent ICLR 2020 [Internet]. 2019 Apr; Available from: http://arxiv.org/abs/1904.00962
43. LeGrand S, Scheinberg A, Tillack AF, Thavappiragasam M, Vermaas JV, Agarwal R, et al. GPU-Accelerated Drug Discovery with Docking on the Summit Supercomputer. In: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics [Internet]. New York, NY, USA: ACM; 2020. p. 1–10. Available from: https://dl.acm.org/doi/10.1145/3388440.3412472
44. Glaser J, Vermaas JV, Rogers DM, Larkin J, LeGrand S, Boehm S, et al. High-throughput virtual laboratory for drug discovery using massive datasets. Int J High Perform Comput Appl. 2021;35(5):452–68.*

The authors demonstrated extremely fast screening of massive chemical databases for COVID-19 drug discovery, docking over one billion compounds to SARS-CoV-2 Mpro and PLpro protein structures at an unprecedented speed of >19,000 compounds docked per second.

45. Wójcikowski M, Ballester PJ, Siedlecki P. Performance of machine-learning scoring functions in structure-based virtual screening. Sci Rep. 2017 Apr;7(1):46710.
46. Ballester PJ, Mitchell JBO. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. Bioinformatics. 2010;26(9):1169–75.
47. Hsu DJ, Lu H, Kashi A, Matheson M, Gounley J, Wang F, et al. TwoFold: Highly accurate structure and affinity prediction for protein-ligand complexes from sequences. Int J High Perform Comput Appl. 2023;37(6):666–82.
48. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583–9.
49. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988 Feb;28(1):31–6.
50. Jacot A, Gabriel F, Hongler C. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. Adv Neural Inf Process Syst. 2018 Jun;2018-Decem[5]:8571–80.
51. Henzler-Wildman K, Kern D. Dynamic personalities of proteins. Nature. 2007 Dec;450(7172):964–72.
52. Pan AC, Borhani DW, Dror RO, Shaw DE. Molecular determinants of drug–receptor binding kinetics. Drug Discov Today. 2013 Jul;18(13–14):667–73.
53. Shan Y, Kim ET, Eastwood MP, Dror RO, Seeliger MA, Shaw DE. How Does a Drug Molecule Find Its Target Binding Site? J Am Chem Soc. 2011 Jun;133(24):9181–3.
54. Pan AC, Xu H, Palpant T, Shaw DE. Quantitative Characterization of the Binding and Unbinding of Millimolar Drug Fragments with Molecular Dynamics Simulations. J Chem Theory Comput. 2017;13(7):3372–7.
55. Gapsys V, Pérez-Benito L, Aldeghi M, Seeliger D, van Vlijmen H, Tresadern G, et al. Large scale relative protein ligand binding affinities using non-equilibrium alchemy. Chem Sci. 2020;11(4):1140–52.
56. Chodera JD, Mobley DL, Shirts MR, Dixon RW, Branson K, Pande VS. Alchemical free energy methods for drug discovery: progress and challenges. Curr Opin Struct Biol. 2011 Apr;21(2):150–60.
57. Páll S, Abraham MJ, Kutzner C, Hess B, Lindahl E. Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS. In: Lecture Notes in Computer Science [including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics] [Internet]. Springer Verlag; 2015. p. 3–27. Available from: https://link.springer.com/chapter/10.1007/978-3-319-15976-8_1 http://link.springer.com/10.1007/978-3-319-15976-8_1
58. Allec SI, Sun Y, Sun J, Chang CEA, Wong BM. Heterogeneous CPU+GPU-Enabled Simulations for DFTB Molecular Dynamics of Large Chemical and Biological Systems. J Chem Theory Comput. 2019;15(5):2807–15.
59. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, et al. The Amber biomolecular simulation programs. J Comput Chem. 2005 Dec;26(16):1668–88.
60. Phillips JC, Hardy DJ, Maia JDC, Stone JE, Ribeiro JV, Bernardi RC, et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. J Chem Phys. 2020 Jul;153(4):044130.
61. Gapsys V, Hahn DF, Tresadern G, Mobley DL, Rampp M, De Groot BL. Pre-Exascale Computing of Protein-Ligand Binding Free Energies with Open Source Software for Drug Design. J Chem Inf Model. 2022;62(5):1172–7.
62. Schindler CEM, Baumann H, Blum A, Böse D, Buchstaller HP, Burgdorf L, et al. Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects. J Chem Inf Model. 2020 Nov;60(11):5457–74.
63. Li Z, Wu C, Li Y, Liu R, Lu K, Wang R, et al. Free energy perturbation–based large-scale virtual screening for effective drug discovery against COVID-19. Int J High Perform Comput

Appl. 2023;37(1):45–57.*

The authors implemented a scalable workflow to perform binding free energy calculations for the discovery of antiviral drugs targeting SARS-CoV-2 Mpro and TMPRSS2 proteins, able to automatically handle hundreds of thousands of molecular dynamics simulation tasks. The authors state that they were able to identify at least one inhibitor that showed promising outcomes in subsequent clinical trials.

64. Li Z, Li X, Huang YY, Wu Y, Liu R, Zhou L, et al. Identify potent SARS-CoV-2 main protease inhibitors via accelerated free energy perturbation-based virtual screening of existing drugs. Proc Natl Acad Sci. 2020 Nov;117(44):27381–7.
65. Zimmerman MI, Porter JR, Ward MD, Singh S, Vithani N, Meller A, et al. SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. Nat Chem. 2021 Jul;13(7):651–9.
66. Boby ML, Fearon D, Ferla M, Filep M, Koekemoer L, Robinson MC, et al. Open science discovery of potent noncovalent SARS-CoV-2 main protease inhibitors. Science. 2023 Nov;382(6671):eabo7201.
67. Ginex T, Vázquez J, Estarellas C, Luque FJ. Quantum mechanical-based strategies in drug discovery: Finding the pace to new challenges in drug design. Curr Opin Struct Biol. 2024 Aug;87:102870.
68. [Lv] WL, Arnesano F, Carloni P, Natile G, Rossetti G. Effect of in vivo post-translational modifications of the HMGB1 protein upon binding to platinated DNA: a molecular simulation study. Nucleic Acids Res. 2018 Dec;46(22):11687–97.
69. Kar RK. Benefits of hybrid QM/MM over traditional classical mechanics in pharmaceutical systems. Drug Discov Today. 2023 Jan 1;28(1):103374.
70. Bolnykh V, Olsen JMH, Meloni S, Bircher MP, Ippoliti E, Carloni P, et al. Extreme Scalability of DFT-Based QM/MM MD Simulations Using MiMiC. J Chem Theory Comput. 2019 Oct;15(10):5601–13.
71. Rossetti G, Mandelli D. How exascale computing can shape drug design: A perspective from multiscale QM/MM molecular dynamics simulations and machine learning-aided enhanced sampling algorithms. Curr Opin Struct Biol. 2024 Jun;86:102814.
72. Raghavan B, Paulikat M, Ahmad K, Callea L, Rizzi A, Ippoliti E, et al. Drug Design in the Exascale Era: A Perspective from Massively Parallel QM/MM Simulations. J Chem Inf Model. 2023 Jun;63(12):3647–58.
73. Antalík A, Levy A, Kvedaravičiūtė S, Johnson SK, Carrasco-Busturia D, Raghavan B, et al. MiMiC: A High-Performance Framework for Multiscale Molecular Dynamics Simulations. 2024 Mar;1–19.
74. Raghavan B, Vivo MD, Carloni P. Metal Coordination and Enzymatic Reaction of the Glioma-Target R132H Isocitrate Dehydrogenase 1 : Insights by Molecular Simulations. ChemRxiv. :1–23.*

This paper demonstrates how multiscale QM/MM MD simulations, enabled by extremely scalable software, can play an invaluable role in structure based drug design campaigns by providing high-quality protein structures for subsequent docking campaigns.

75. Scarpino A, Ferenczy GG, Keserű GM. Comparative Evaluation of Covalent Docking Tools. J Chem Inf Model. 2018 Jul;58(7):1441–58.
76. Carrasco-Busturia D, Ippoliti E, Meloni S, Rothlisberger U, Olsen JMH. Multiscale biomolecular simulations in the exascale era. Curr Opin Struct Biol. 2024 Jun;86:102821.
77. Alexander F, Almgren A, Bell J, Bhattacharjee A, Chen J, Colella P, et al. Exascale applications: Skin in the game. Philos Trans R Soc Math Phys Eng Sci. 2020;378(2166].
78. Agosta G, Cattaneo D, Fornaciari W, Galimberti A, Massari G, Reghenzani F, et al. Textarossa: Towards EXtreme scale Technologies and Accelerators for euROhpc hw/Sw Supercomputing Applications for exascale. Proc - 2021 24th Euromicro Conf Digit Syst Des DSD 2021. 2021;286–94.
79. Palermo G, Accordi G, Gadioli D, Vitali E, Silvano C, Guindani B, et al. Tunable and Portable Extreme-Scale Drug Discovery Platform at Exascale: the LIGATE Approach. 2023 Apr; Available from: http://arxiv.org/abs/2304.09953%0Ahttp://dx.doi.org/10.1145/3587135.3592172 http://arxiv.org/abs/2304.09953 http://dx.doi.org/10.1145/3587135.3592172
80. Palermo G, Accordi G, Gadioli D, Zhang Y, Vitali E, Guindani B, et al. LIGATE-LIgand Generator and portable drug discovery platform at Exascale. Proc 21st ACM Int Conf Comput Front 2024 Workshop Spec Sess CF 2024 Companion. 2024;107–9.
81. Zeng X, Wang F, Luo Y, Kang S gu, Tang J, Lightstone FC, et al. Deep generative molecular design reshapes drug discovery. Cell Rep Med. 2022 Dec;3(12):100794.
82. Gavini V, Baroni S, Blum V, Bowler DR, Buccheri A, Chelikowsky JR, et al. Roadmap on electronic structure codes in the exascale era. Model Simul Mater Sci Eng. 2023 Sep;31(6):063301.
83. Yokelson D, Tkachenko NV, Robey R, Li YW, Dub PA. Performance Analysis of CP2K Code for Ab Initio Molecular Dynamics on CPUs and GPUs. J Chem Inf Model [Internet]. 2022 Apr; Available from: https://pubs.acs.org/doi/10.1021/acs.jcim.1c01538
84. Carnimeo I, Affinito F, Baroni S, Baseggio O, Bellentani L, Bertossa R, et al. Quantum ESPRESSO: One Further Step toward the Exascale. J Chem Theory Comput. 2023;
85. Das S, Motamarri P, Subramanian V, Rogers DM, Gavini V. DFT-FE 1.0: A massively parallel hybrid CPU-GPU density functional theory code using finite-element discretization. 2022 Mar;1–55.
86. Ufimtsev IS, Martinez TJ. Quantum Chemistry on Graphical Processing Units. 3. Analytical Energy Gradients, Geometry Optimization, and First Principles Molecular Dynamics. J Chem Theory Comput. 2009 Oct;5(10):2619–28.
87. Cruzeiro VWD, Wang Y, Pieri E, Hohenstein EG, Martínez TJ. TeraChem protocol buffers [TCPB): Accelerating QM and QM/MM simulations with a client–server model. J Chem Phys. 2023 Jan;158(4):044801.
88. Manathunga M, Aktulga HM, Götz AW, Merz KM. Quantum Mechanics/Molecular Mechanics Simulations on NVIDIA and AMD Graphics Processing Units. J Chem Inf Model. 2023 Feb;63(3):711–7.
89. Rizzi A, Carloni P, Parrinello M. Targeted Free Energy Perturbation Revisited: Accurate Free Energies from Mapped Reference Potentials. J Phys Chem Lett. 2021 Oct;12(39):9449–54.
90. Rizzi A, Carloni P, Parrinello M. Free energies at QM accuracy from force fields via

multimap targeted estimation. Proc Natl Acad Sci. 2023 Nov;120(46):1–10.
91. Kozinsky B, Musaelian A, Johansson A, Batzner S. Scaling the Leading Accuracy of Deep Equivariant Models to Biomolecular Simulations of Realistic Size. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (Internet). New York, NY, USA: Association for Computing Machinery; 2023. (SC '23). Available from: https://doi.org/10.1145/3581784.3627041
92. Jia W, Wang H, Chen M, Lu D, Lin L, Car R, et al. Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE Press; 2020. (SC '20).

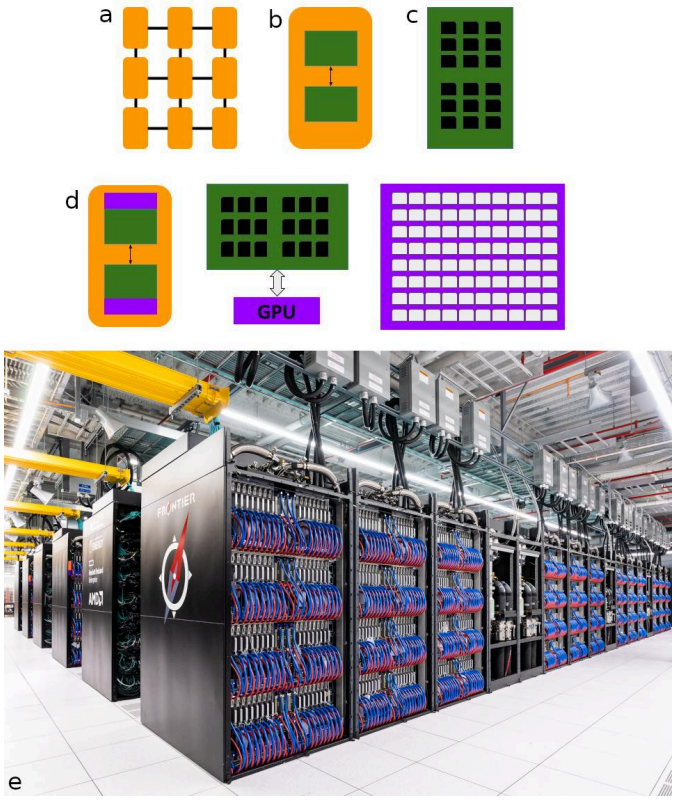

Figure 1: (a) A modern supercomputer can be schematically represented as a set of interconnected computing nodes. (b) Each compute node contains several CPUs mounted on the same board. (c) Modern CPUs nowadays host several (4, 8, 12 or more) general purpose cores residing on the same die. (d) Current exascale supercomputers make use of heterogeneous nodes,

where CPUs are coupled to one or more GPUs, each hosting hundreds of compute units. (e) The FRONTIER exascale supercomputer at the Oak Ridge National Laboratories counts a total of 9,408 compute nodes. Each compute node contains one 64-core CPU, with access to 512 Gb of RAM, and eight GPUs. Each GPU contains 110 compute units and has access to 64 GB of high-bandwidth memory. (The image of the FRONTIER supercomputer was originally posted to Flickr by OLCF at https://flickr.com/photos/151938121@N02/52117623843. It was reviewed on 13 June 2022 by FlickreviewR 2 and was confirmed to be licensed under the terms of the cc-by-2.0 (https://creativecommons.org/licenses/by/2.0/))

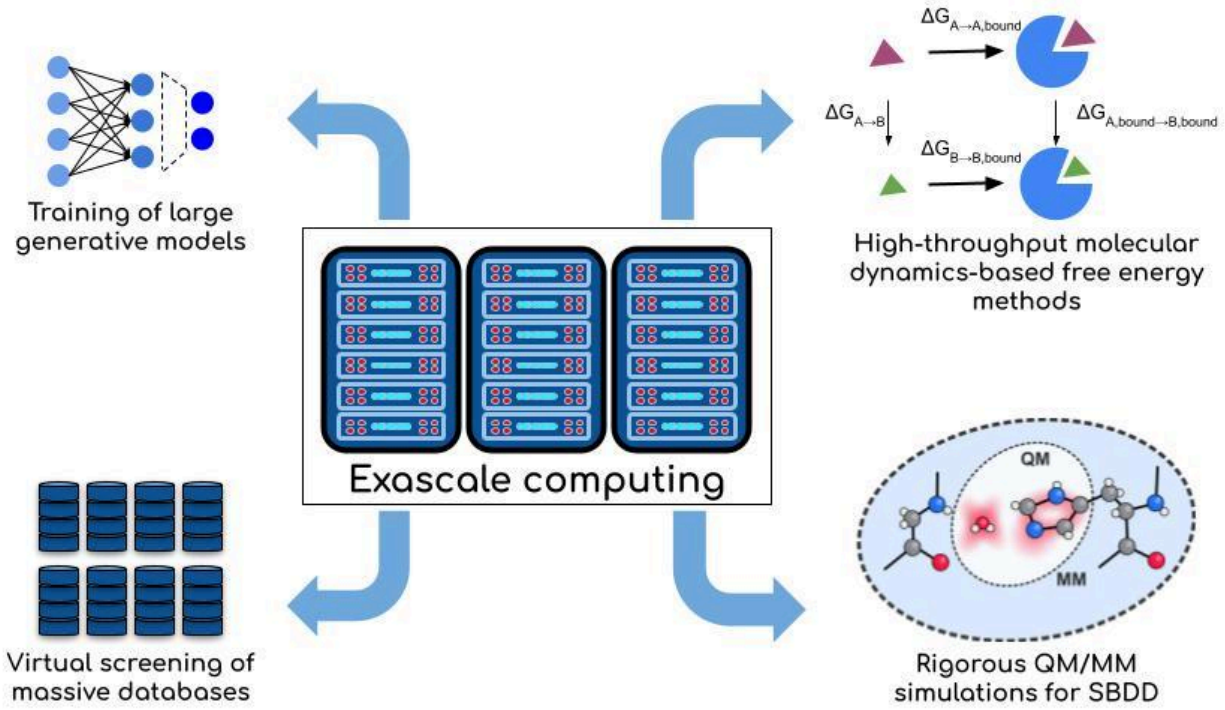

Figure 2: The diagram illustrates the different HPC-oriented CADD approaches for the design of small molecule binders that can take advantage of exascale computing. (The QM/MM schematics (bottom right) was originally posted at https://www.cp2k.org/_detail/exercises:2016_summer_school:qmmmcartoon.png?id=events%3A2016_summer_school%3Aqmmm. It is licensed under the following license:

Table 1: The table reports the main achievements of selected works that leveraged (pre)exascale supercomputers to push the boundaries of different computational methods in drug design and discovery.

Original papers	**Computational Resources**	**Category**	**Main Achievement**
Jacobs et al.	4,320 GPU-nodes of the Sierra supercomputer at LLNL	Machine Learning	Record time of 23 minutes to train a generative model on a dataset of >1.6 billion compounds
Blanchard et al.	4,032 GPU-nodes of the Summit supercomputer at OLCF	Machine Learning	Record time of few hours to pre-train a large language model on a dataset of 9.6 billion molecules

Glaser et al.	4,602 GPU-nodes of the Summit supercomputer at OLCF	Docking	Record time for molecular docking 1.37 billion compounds to SARS-CoV-2 proteins, reaching an average rate of 19,028 compounds docked per second
Hsu et al.	128 GPU-nodes of the Frontier exascale supercomputer at OLCF 4,056 GPU-nodes of the Summit supercomputer at OLCF	Machine Learning	Train the first deep-learning model for protein-ligand structure prediction from sequence. Train and validate an infinitely wide deep neural network for binding affinity predictions, solving a 1.15M linear system within 1.5 minutes

Li et al.	75,000 nodes of the new generation Tianhe supercomputer	Molecular Dynamics	Complete a free energy perturbation-based virtual screening of ~12,000 ligand-receptor pairs within six days
Ragavan et al.	1,746 CPU-nodes of the JUWELS Cluster supercomputer at JSC	QM/MM Molecular Dynamics	The work reports unprecedented strong scaling in a QM/MM MD simulation of a pharmacologically relevant enzyme